Discrete Applied Mathematics 157 (2009) 2244–2256

Contents lists available at ScienceDirect

Discrete Applied Mathematics

journal homepage: www.elsevier.com/locate/dam

Path finding approaches and metabolic pathways

F.J. Planes [a,b], J.E. Beasley [a,*]

[a] Department of Mathematical Sciences, Brunel University, Uxbridge, UB8 3PH, UK

[b] CEIT and TECNUN, University of Navarra, Manuel de Lardizabal 15, 20018 San Sebastian, Spain

ARTICLE INFO

Article history:
Received 31 October 2006
Received in revised form 25 June 2007
Accepted 11 June 2008
Available online 20 September 2008

Keywords:
Metabolic pathways
Path finding
Shortest path
Systems biology

ABSTRACT

A metabolic pathway is a set of enzyme catalysed biochemical reactions by which a living organism transforms an initial (source) compound into a final (target) compound. Path finding approaches to metabolic pathways adopt a graph theory approach to the problem of determining the reactions an organism might use to transform a source compound into a target compound. In this paper, the effectiveness of using compound node connectivities in a path finding approach is examined. An approach to path finding based upon integer programming is also presented.

1. Introduction

The complex world of cellular metabolism is commonly organised into metabolic pathways. A number of metabolic pathways, which are usually defined as a set of enzyme catalysed biochemical reactions, by which a living organism transforms an initial source compound into a final target compound, have been elucidated via experiments on different model organisms (e.g. Nelson and Cox [19]).

The problem of determining in a computational (algorithmic) fashion these metabolic pathways, given a database of biochemical reactions and compounds associated with a particular organism or cell, has attracted increased interest in recent years. We later review such work below. First, however, we present a basic introduction to metabolic pathways and path finding approaches.

1.1. Metabolic pathways and path finding approaches

Path finding approaches to metabolic pathways view the set of possible reactions (and their associated compounds) that can appear in a particular living organism as a directed graph (*metabolic network*). The usual form that this directed graph takes is that the nodes (vertices) of the graph represent reactions and compounds. Fig. 1 shows an example metabolic network comprising 6 reactions (labelled R1 to R6 respectively) and 8 compounds (labelled C1 to C8 respectively). Reaction R3 (for example) converts C5 into C4, C6 and C7, as indicated by the arcs of the graph. Note here that the graph shown in Fig. 1 is bipartite, the set of reaction nodes {R1 to R6} and the set of compound nodes {C1 to C8} have arcs between them, but there are no reaction-to-reaction, or compound-to-compound, arcs.

In this paper, we have had to assume some knowledge of graph theory on the part of the reader. Readers unfamiliar with graph theory can consult one of the many basic textbooks available should they have difficulty with some of the graph theory concepts used below. A good overview of the application of graph theory with respect to problems that arise in cell biology (including metabolic pathways) has recently been given by Aittokallio and Schwikowski [1].

* Corresponding author.
E-mail addresses: Francisco.Planes@brunel.ac.uk (F.J. Planes), john.beasley@brunel.ac.uk (J.E. Beasley).

0166-218X/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.dam.2008.06.035

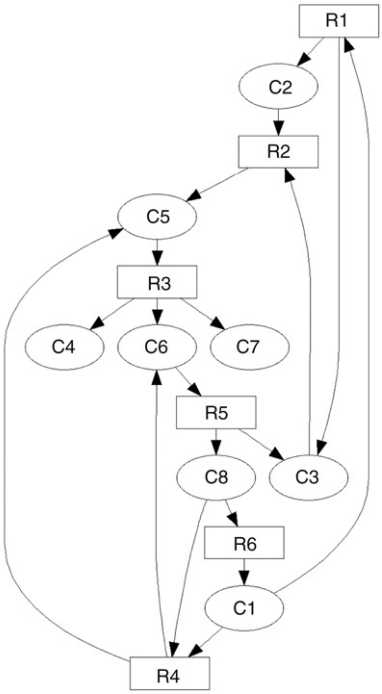

Fig. 1. An example metabolic network.

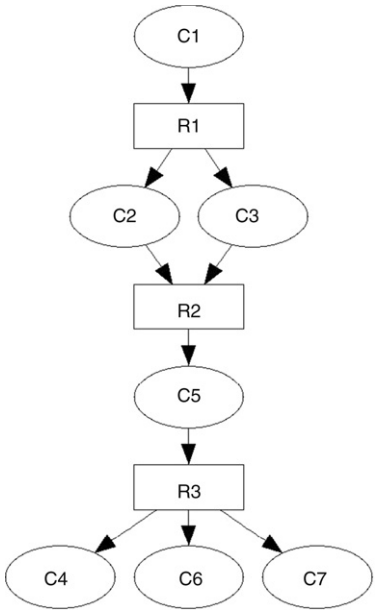

Fig. 2. One possible subgraph (pathway) for converting C1 into C7.

Suppose we are considering the metabolic pathway that converts C1 into C7. This pathway will be a subgraph of this entire metabolic network (which is a directed graph). As one might expect, within a directed graph representation of a metabolic network, there may be more than one such subgraph. Two subgraphs of Fig. 1 that convert C1 into C7 are shown in Figs. 2 and 3. Both of these subgraphs can essentially be regarded as feasible pathways from a biochemical viewpoint. Note here that the subgraphs shown in Figs. 2 and 3 do have the properties that:

- for each reaction node included in the subgraph all compound nodes associated with that reaction node (either as an input compound or as an output compound) also appear in the subgraph
- there is a (directed) path in the subgraph from the source node (compound C1) to the target node (compound C7).

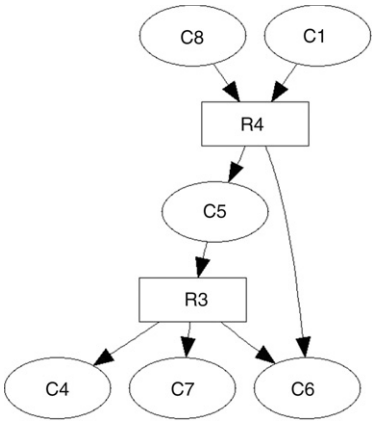

Fig. 3. Another possible subgraph (pathway) for converting C1 into C7.

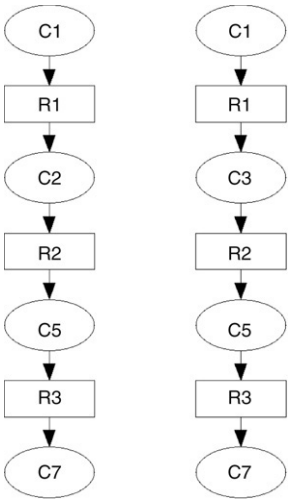

Fig. 4. Two metabolic paths in the metabolic pathway shown in Fig. 2.

Experimental findings on different organisms indicate that (despite the presence of many possible biochemically feasible pathways) organisms often have a single "preferred" pathway, in order to convert a source compound into a target compound. Such findings are reflected in, for example, standard textbooks such as Nelson and Cox [19], or online resources such as BioCyc (http://biocyc.org/). Here, the terminology is not unique, and different authors describe these pathways using different terms, e.g. annotated pathway [4,5]; consensus pathway [2]; experimentally elucidated pathway [10]; experimentally determined pathway [3]; reference pathway (Kegg Pathway Database, http://www.genome.ad.jp/kegg/pathway.html). Of course, it is important to note that from the physiological viewpoint, metabolic pathways do not operate in isolation, and within an organism many different pathways work together to produce an overall global flux (reaction/compound) distribution.

Suppose, for illustration, that the observed (experimentally elucidated/determined) metabolic pathway which converts C1 into C7 is the subgraph shown in Fig. 2. Fig. 2 therefore is the *metabolic pathway*. In this pathway C1 is the source compound (sometimes referred to as a required/desired reactant, Mavrovouniotis et al. [14]) and C7 is the target compound (sometimes referred to as a required product, Mavrovouniotis et al. [14]). Compounds C2, C3, and C5 are intermediate compounds, which are produced as well as consumed; compounds C4 and C6 are produced, but not consumed.

Path finding approaches to metabolic pathways focus on a directed path (containing no cycles) from the source compound (here C1) to the target compound (here C7) in the metabolic pathway. We refer to this directed path as the *metabolic path* for the pathway. Of course this path may not be unique, in particular when the pathway is branched. Here, for example, we have in Fig. 2 the two paths shown in Fig. 4.

Note, as in Figs. 2 and 4, the difference between a metabolic pathway and a metabolic path. The metabolic pathway (Fig. 2) contains all the reactions and compounds involved in the pathway. The metabolic path is a directed path from the source compound to the target compound in the metabolic pathway, and may (as in both the metabolic paths seen in Fig. 4) contain only a subset of intermediate reactions/compounds.

For an unknown metabolic pathway (but with a given source and target compound), the key assumption behind path finding approaches is that finding directed paths between the source compound and the target compound in the entire metabolic network will give insight into the intermediate reactions/compounds used in the metabolic pathway between the source/target.

In terms of the metabolic pathway we need only focus on the reactions involved in the metabolic path (since for each reaction we know the set of compounds involved). Both of the metabolic paths shown in Fig. 4 involve reactions R1, R2 and R3. Since, from Fig. 2, these are the only reactions involved in the metabolic pathway then, for this example, knowledge of either of the metabolic paths shown in Fig. 4 would give us complete insight into the underlying metabolic pathway, as shown in Fig. 2. Note here, that with respect to the stoichiometry of the metabolic pathway shown in Fig. 2 then, once the set of reactions involved in the pathway are known, it is a relatively simple matter (albeit possibly involving some decisions as to which compounds should be balanced with respect to production and consumption), to deduce the stoichiometry of the pathway. With respect to deciding which compounds might be balanced with respect to production and consumption, we would note that the intermediate compounds in the metabolic path are themselves often balanced compounds.

Early approaches to path finding given in the literature consisted of simple enumeration of paths (or pathways), from the source node to the target node in the directed graph representation of the metabolic network. However the work of Küffner et al. [13], who showed that there were some 500,000 paths from glucose to pyruvate, illustrated that a more sophisticated approach was needed. Accordingly the focus of path finding work has moved to defining a suitable distance metric on the directed graph representation of the metabolic network and finding the shortest path from the source node to the target node. Often, approaches in the literature move beyond considering just the shortest path to consider the k-shortest paths (for small values of k). For readers unfamiliar with the concept of k-shortest paths $k = 1$ corresponds to the shortest path; $k = 2$ corresponds to the second shortest path; $k = 3$ to the third shortest path; etc.

In the next section, we present a literature review of relevant work dealing with path finding in metabolic networks. Other approaches to metabolic pathways, for example flux balance analysis, extreme pathways and elementary flux modes will not be discussed in any detail here, but relevant work can be found in Klamt and Stelling [11]; Papin et al. [20]; Schilling et al. [24,25]; Schuster et al. [26] and Stelling et al. [30].

1.2. Path finding literature review

Early work presented in the literature is Seressiotis and Bailey [27,28] and Mavrovouniotis [14–17]. Seressiotis and Bailey [27,28] developed an algorithm based on artificial intelligence concepts, so as to find pathways transforming a source compound into a target compound. They dealt with only a relatively small reaction/compound database (70 reactions, 100 compounds). Mavrovouniotis [14–17] developed an algorithm which also generated pathways. However, it required a specification of compound status (e.g. which compounds must be consumed in the pathway). That work [17] dealt with a database of 250 reactions and 400 compounds.

Küffner et al. [13] described the database of reactions/compounds as a Petri Net (bipartite graph), where there are two type of nodes, places (compounds) and transitions (reactions). The edges connect places (input compounds) with transitions and transitions with places (output compounds). Paths/pathways are formed via a "firing rule". The solution approach adopted was a branch and bound algorithm. They found that there were over 500,000 paths from glucose to pyruvate. Their firing rule reduced the total number of paths to approximately 80,000. This was reduced to 170 pathways by imposing further restrictions.

Two themes present in Küffner et al. [13], namely establishing all possible paths between two compounds, and using Petri Nets in metabolic networks, are still active themes in the literature. For example with regard to establishing all possible paths between two compounds Hatzimanikatis et al. [8] examined paths from chorismate and found that there are more than 350,000 paths to tyrosine; 75,000 to phenylalanine; but only 13 to tryptophan. With regard to using Petri Nets in metabolic networks, Zevedei-Oancea and Schuster [31] provide a good overview; Koch et al. [12] applied them to the pathway associated with the breakdown of sucrose in potato tubers; Simao et al. [29] applied them to the pathway from chorismate to tryptophan (cf Hatzimanikatis et al. [8]).

Because of the large number of possible paths identified by Küffner et al. [13], much subsequent work reported in the literature has focused on enumerating just a small number of paths. We describe these path finding approaches below.

Arita [2] proposed the use of a k-shortest path algorithm to find paths in metabolic networks, where compounds are represented at the atomic level. They applied their approach to a number of example pathways, where "shortest" is interpreted as minimising the number of arcs (reactions) involved in the path. For Glycolysis (regarded as a path from glucose to pyruvate) they reported that they find some, but not all, of the compounds appearing in that path. They noted that one advantage of their approach is the enumeration of multiple (or in the limit, all) paths.

McShan et al. [18] considered metabolic pathways in terms of a biochemical state-space: compounds define the states and reactions define the state-transitions. The state-space, compounds, are defined as a vector $x = (x_1, x_2, \ldots, x_n)$ of 145 chemical descriptors. The state-transitions, reactions, are considered as transitions between states. Each reaction is simplified to only one input and output compound, avoiding side compounds. The cost of the transitions is defined as the Manhattan distance of the Δx vector, Δx being defined as the difference between the x vectors belonging to the input and output compounds. The problem of finding metabolic paths is viewed as searching for a path from an initial state to a destination state, through a series of transitions. An algorithm (A* search) to minimise the cost of the transitions was

applied to find metabolic paths. They reported that they found A* search to be more efficient than other search techniques they examined, such as breadth-first or depth-first search.

Dooms et al. [6] proposed the use of constraint programming to find constrained paths in metabolic networks. They cite the Ph.D. thesis of Croes (work later reported in [4,5]) and, although their wording is imprecise, it does appears that all compound nodes in their approach were assigned a weight proportional to their degree of connectivity (number of reactions in which the compound participates), as in [4,5]. One limitation is that in their work, they need to know some of the reactions participating in the metabolic path that represents the metabolic pathway. In addition, they note that their approach cannot guarantee to find the optimal constrained shortest path.

In Rahman et al. [21], by comparison to other approaches, there are no reactions nodes in the directed graph, only compound nodes. Edges between any two compounds are assigned according to their structural similarity. A breadth-first search algorithm was applied to compute the k-shortest paths between an initial source compound and a final target compound. In related work Rahman and Schomburg [22], used a k-shortest path approach to identify "load points" and "choke points". They defined a load value for each compound based on the ratio of the number of k-shortest paths passing through it, and the number of links associated with the compound. They defined a choke value for each compound, based on the number of k-shortest paths passing through it and the load value. Results from their approach were presented for two related bacteria.

Croes et al. [4,5] presented a path finding approach that utilises connectivity. They define connectivity for a compound to be the number of reactions (in the reaction database) in which the compound participates (either as an input compound or as an output compound). They define connectivity for a reaction node to be one. Node connectivities are then taken as the distance metric to be minimised when finding shortest paths. They use a depth-first backtracking (tree search) algorithm to find the k-shortest paths ($k = 1, 2, 3, 4, 5$) *not between a source compound and a target compound, but between a source reaction and a target reaction.* Their view of a metabolic pathway as being between a source reaction and a target reaction is not usual in the literature. They systematically (and numerically) compare the k-shortest paths they find with a number of metabolic paths. Their approach, which appears to be the most effective of all path finding approaches presented to date in the literature, is based on the observation that many of the intermediate compounds in a metabolic path appear to have low connectivity.

In a recent paper, we (Beasley and Planes [3]) have presented an approach to metabolic pathways that is based on integer programming, though not involving path finding. That approach focuses not on the metabolic path, but on the metabolic pathway.

In the next section we outline the contribution of this paper.

1.3. Contribution

In this paper, the effectiveness of using compound node connectivities in conjunction with k-shortest path calculations, as initially proposed by Croes et al. [4,5], is examined in ten *E.Coli.* metabolic pathways. As noted above, the work of Croes et al. [4,5] is unusual, in that they only consider paths from a source reaction to a target reaction (which we denote as the R–R case). In this paper, we also consider paths from a source compound to a target compound (which we denote as the C–C case). Moreover, we present results for higher values of k (up to $k = 10$) than Croes et al. [4,5] (they considered up to $k = 5$), so as to see the benefit of increasing the number of shortest paths considered.

It is clear from our reading of the various papers discussed above, that authors have taken an approach to calculating paths based upon algorithms such as breadth-first and depth-first search. Such algorithms, although relatively easy to code, are often computationally ineffective (especially for paths that involve many nodes). As such a detailed examination of papers discussed above often reveals some choice being made so as to limit computational effort. For example:

- Dooms et al. [6], constraint programming, impose a limit of the size of the directed graph
- Croes et al. [5], depth-first search, impose an upper limit on the number of nodes in the path and the total length of the path.

Such choices, whilst being necessary for computational reasons, do mean that the paths found may not (in fact) be optimal, i.e. there may exist shorter paths that have been missed because of these heuristic choices.

In addition, algorithms such as breadth-first and depth-first search do not produce paths in increasing distance order, i.e. they do not first find the ($k = 1$) shortest path; then the ($k = 2$) second shortest path; etc. Rather, the entire search algorithm must be allowed to finish enumerating paths in the directed graph (many of which will be irrelevant), before all of the k-shortest paths are known.

In this paper, we are concerned with finding k-shortest paths between a source node and a target node in a directed graph (metabolic network), where in each path no node appears more than once. In order to do this we present below an integer programming approach that produces paths in increasing distance order. Although other approaches to finding k-shortest paths are available, e.g. see [7], we believe using integer programming does have advantages over previous approaches used for calculating k-shortest paths in the metabolic pathway literature, in terms of:

- producing paths in increasing distance order; and
- guaranteeing that the paths found will be optimal.

Our integer programming approach to calculating k-shortest paths is presented in the next section. As stated above, the key assumption behind path finding approaches is that finding directed paths between the source compound and the target compound in the entire metabolic network will give insight into the metabolic pathway between these compounds. As such, we believe that refining path finding approaches, as is done in this paper, is of value as they may contribute both to elucidating unknown pathways and to providing more insight into existing pathways.

2. Integer programming approach

2.1. Formulation

In our approach we have a database of R reactions (where each reaction has a specified direction so a reversible reaction contributes two different reactions to the total number R), which collectively involve C different compounds. Let m_{cr} have the value 1 if compound c is an input compound for reaction r, 0 otherwise. Let d_{rc} have the value 1 if compound c is an output compound from reaction r, 0 otherwise. Let $W_c = \sum_{r=1}^{R} \max(m_{cr}, d_{rc})$ be the connectivity of compound c, i.e. the number of reactions in which the compound appears in the database of reactions. Assuming no compound is both input and output from the same reaction W_c can also be viewed as the sum of the in-degree and out-degree of compound c in the directed graph representation.

Suppose we are seeking the shortest path from a source node S to a target node T in our directed graph representation where, for ease of exposition, we assume below that S and T are compound nodes. Amending the formulation given below if S and T are reaction nodes is easily done.

Variables

We need to decide the arcs involved in the metabolic path, so our zero-one (binary, integer) variables are:

- $u_{cr} = 1$ if the arc from compound node c to reaction node r is in the metabolic path; 0 otherwise
- $v_{rc} = 1$ if the arc from reaction node r to compound node c is in the metabolic path; 0 otherwise.

If $m_{cr} = 0$, i.e. the arc does not exist, then we fix u_{cr} to 0; similarly if $d_{rc} = 0$ we fix v_{rc} to 0. This enables us to present the constraints below in a simplified form.

Constraints

The constraints are:

$$\sum_{r=1}^{R} u_{Sr} = \sum_{r=1}^{R} v_{rT} = 1 \tag{1}$$

$$\sum_{r=1}^{R} v_{rS} = \sum_{r=1}^{R} u_{Tr} = 0. \tag{2}$$

Eq. (1) ensures that one arc leaves S and one arc enters T. Eq. (2) that no arc enters S and no arc leaves T.

$$\sum_{c=1}^{C} u_{cr} = \sum_{c=1}^{C} v_{rc} \quad r = 1, \ldots, R \tag{3}$$

$$\sum_{r=1}^{R} v_{rc} = \sum_{r=1}^{R} u_{cr} \quad c = 1, \ldots, C \; c \neq S, T. \tag{4}$$

Eq. (3) ensures that the number of arcs entering a reaction node is equal to the number leaving. Eq. (4) fulfils the same condition for compound nodes.

$$\sum_{c=1}^{C} u_{cr} \leq 1 \quad r = 1, \ldots, R \tag{5}$$

$$\sum_{r=1}^{R} v_{rc} \leq 1 \quad c = 1, \ldots, C \; c \neq S, T. \tag{6}$$

Eqs. (5) and (6) ensure that no reaction/compound node is revisited in the path.

We need constraints to prevent cycles appearing. Referring back to Fig. 1, if we are seeking a path from C1 to C7 then Fig. 5, where we do have a path from C1 to C7 but also a cycle R4 $\rightarrow$ C6 $\rightarrow$ R5 $\rightarrow$ C8 $\rightarrow$ R4, is a valid solution to the constraints presented so far above.

Note here on a technical issue, that if we are seeking just the ($k = 1$) shortest path, then cycles will not appear as the distance metric (W_c) we use is non-negative. However, because we intend to use our formulation to find k-shortest paths (for $k \geq 2$) cycles may appear.

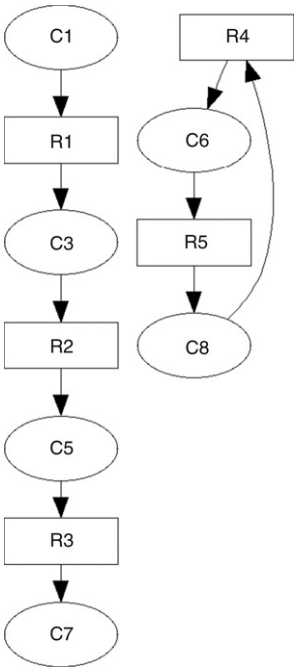

Fig. 5. A cycle.

Cycle elimination constraints are standard in the literature. As an illustration, for the cycle with four arcs (R4 $\rightarrow$ C6, C6 $\rightarrow$ R5, R5 $\rightarrow$ C8, C8 $\rightarrow$ R4) shown in Fig. 5 the cycle elimination constraint is $(v_{R4,C6} + u_{C6,R5} + v_{R5,C8} + u_{C8,R4}) \leq 3$. In general the constraint to eliminate a cycle is: (sum of the v_{rc} and u_{cr} variables for arcs appearing in the cycle) $\leq$ (number of arcs in the cycle -1). Computationally cycle elimination constraints are added as and when cycles appear in solutions (identifying a cycle in a directed graph is algorithmically an easy task). Adding constraints to eliminate cycles, as and when they appear, is standard computational practice (since adding constraints to prevent any cycle at all appearing entails adding a very large number of constraints). After adding a cycle elimination constraint we resolve the problem, this process of adding cycle elimination constraints and resolving being repeated until no cycles exist in the solution.

Other authors dealing with path finding in metabolic pathways (e.g. McShan et al. [18], Croes et al. [4,5]) constrain paths so as to exclude a reaction and its reverse. This is easily done within our integer programming approach. Let B be the set $\{(\alpha, \beta)|$reaction node α and reaction node β are the reverse of each other, $\alpha < \beta\}$. If a reaction r is in the path then it must be true that $\sum_{c=1}^{C} u_{cr} = 1$. So, to prevent reactions α and β from both being in the path we have:

$$\sum_{c=1}^{C} u_{c\alpha} + \sum_{c=1}^{C} u_{c\beta} \leq 1 \quad \forall(\alpha, \beta) \in B. \tag{7}$$

Objective

The objective is to minimise the total connectivity of the compounds involved in the path, i.e.

$$\text{minimise } W_S + \sum_{c=1, c\neq S,T}^{C} W_c \sum_{r=1}^{R} (v_{rc} + u_{cr})/2 + W_T \tag{8}$$

where $\sum_{r=1}^{R} (v_{rc} + u_{cr})/2$ will have the value one if compound c is in the path, and the value zero if compound c is not in the path.

2.2. Solution elimination constraints

In order to find the k-shortest path, we need to add further constraints to eliminate the $(k-1)$-shortest paths from the set of solutions. To illustrate this, suppose we are interesting in finding the $(k = 2)$ second shortest path. Let U_{cr}^1 and V_{rc}^1 be the solution for the $(k = 1)$ shortest path. We need to eliminate this shortest path from the set of solutions. To do this we add the following constraint to our formulation:

$$\sum_{c=1}^{C} \sum_{r=1, U_{cr}^1=0}^{R} u_{cr} + \sum_{c=1}^{C} \sum_{r=1, V_{rc}^1=0}^{R} v_{rc} + \sum_{c=1}^{C} \sum_{r=1, U_{cr}^1=1}^{R} (1 - u_{cr}) + \sum_{c=1}^{C} \sum_{r=1, V_{rc}^1=1}^{R} (1 - v_{rc}) \geq 1. \tag{9}$$

This constraint ensures that the Hamming distance between:

- the ($k = 1$) shortest path solution, and
- the path solution found after this constraint is added and the problem resolved

is at least one, which means that there is a difference of at least one arc between the two solutions. Therefore, if we add this constraint to our formulation and resolve, we will find a new path solution, which is different from the previous solution (the 1-shortest path). Because we are minimising (c.f. Eq. (8)) this new path will be the "next best" path in objective function terms—so it must be the 2-shortest path.

In the general case, in order to find the k-shortest path, we have to include $k-1$ solution elimination constraints as below related to the $(k-1)$-shortest paths:

$$\sum_{c=1}^{C}\sum_{r=1,U_{cr}^{K}=0}^{R} u_{cr} + \sum_{c=1}^{C}\sum_{r=1,V_{rc}^{K}=0}^{R} v_{rc} + \sum_{c=1}^{C}\sum_{r=1,U_{cr}^{K}=1}^{R} (1-u_{cr}) + \sum_{c=1}^{C}\sum_{r=1,V_{rc}^{K}=1}^{R} (1-v_{rc}) \geq 1 \quad K = 1,\ldots,k-1 \tag{10}$$

where U_{cr}^{K} and V_{rc}^{K} are the solution for the K-shortest path.

2.3. Overview

The formulation given above for finding the k-shortest metabolic path is a linear integer (zero-one, binary) program. Algorithmically, such programs are solved by linear programming based tree search, which guarantees that the solution found will be optimal. Modern software packages for solving linear integer (zero-one, binary) programs, such as ILOG Cplex [9] which we used, are well developed and highly sophisticated.

Hence, to summarise, we have presented above a formulation for finding k-shortest paths to which standard software can be applied that:

- produces paths in increasing distance order; and
- guarantees that the paths found will be optimal.

3. Results

3.1. Introduction

The approach to finding k-shortest paths given above was applied to a number of well-known pathways frequently encountered in biochemistry texts (e.g. Nelson and Cox [19]). The reaction/compound database used was drawn from Reed et al. [23] and http://systemsbiology.ucsd.edu/organisms/ecoli/ecoli_reactions.html; and the pathways from Kesler et al. [10], Nelson and Cox [19] and http://biocyc.org/ECOLI/. Details as to the pathways considered in this paper can be found in the accompanying supplement.

In order to judge the effectiveness of our path finding approach, we will compare each path found with a single metabolic path associated with each pathway. As noted above in our discussion of Fig. 4 there may be more than one metabolic path associated with a pathway. In this event, we choose (in a structured way as described in detail in the accompanying supplement) from amongst the possible metabolic paths just one path against which to compare our results.

Two different cases were considered: the reaction to reaction (R–R) case, where paths are computed from the first reaction to the last reaction in the pathway (such as was considered in Croes et al. [4,5]); and the compound to compound (C–C) case, where paths are computed from source compound to target compound in the pathway. The computed paths were evaluated using the same criteria as in Croes et al. [4,5]. These criteria, detailed below, essentially measure the degree of correspondence between any computed path and a path (the metabolic path) that represents the metabolic pathway.

In order to compare the computed path and the metabolic path, Croes et al. [4,5] defined the following correspondence values which indicate, numerically, correspondence between the computed path and the metabolic path:

- *True positives (TP):* The total number of reactions and compounds found in the computed path that are also in the metabolic path. The source and target nodes, whether reaction or compound, are not considered.
- *False positives (FP):* The total number of reactions and compounds found in the computed path that are not in the metabolic path.
- *False negatives (FN):* The total number of reactions and compounds found in the metabolic path that are not in the computed path.
- *Sensitivity (Sn):*$=$ TP/(TP + FN), is the fraction of the reactions and compounds in the metabolic path (excluding source and target) that are in the computed path.
- *Positive Predictive Value (PPV):*$=$ TP/(TP + FP), is the fraction of the reactions and compounds in the computed path (excluding source and target) that are in the metabolic path.
- *Accuracy (Ac):*$=$ (Sn + PPV)/2, is the average of the previous two values.

Table 1
Correspondence values for the first ten shortest paths in the Gluconeogenesis pathway in the R–R case

k shortest path, k	True positives (TP)	False positives (FP)	False negatives (FN)	Sensitivity (Sn)	Positive predictive value (PPV)	Accuracy (Ac)
1	5	2	8	0.385	0.714	0.549
2	10	1	3	0.769	0.909	0.839
3	10	1	3	0.769	0.909	0.839
4	3	4	10	0.231	0.429	0.330
5	3	4	10	0.231	0.429	0.330
6	13	0	0	1	1	1
7	2	13	11	0.154	0.133	0.144
8	2	13	11	0.154	0.133	0.144
9	2	13	11	0.154	0.133	0.144
10	10	3	3	0.769	0.769	0.769

Table 2
Correspondence values for the first ten shortest paths in the Gluconeogenesis pathway in the C–C case

k shortest path, k	True positives (TP)	False positives (FP)	False negatives (FN)	Sensitivity (Sn)	Positive predictive value (PPV)	Accuracy (Ac)
1	2	15	13	0.133	0.118	0.125
2	7	2	8	0.467	0.778	0.622
3	0	21	15	0	0	0
4	0	27	15	0	0	0
5	2	23	13	0.133	0.080	0.107
6	2	21	13	0.133	0.087	0.110
7	0	18	15	0	0	0
8	0	23	15	0	0	0
9	2	23	13	0.133	0.080	0.107
10	0	20	15	0	0	0

Sensitivity, positive predictive value, and accuracy are all defined such that higher values represent closer correspondence between the computed path and the metabolic path. If the computed path corresponds exactly to the metabolic path then Sn = PPV = Ac = 1 (equivalently FP = FN = 0).

Typically, as in this paper, the effectiveness of any path finding approach is examined by seeing how well it performs (for example as evaluated by the above correspondence values) with respect to a known metabolic pathway. In other words given the source and target compound, and the entire metabolic network, how well does a particular path finding approach do at discovering the reactions and compounds involved in a known metabolic path or pathway?

Note here, that we have adopted (as detailed above) the same correspondence values as defined in Croes et al. [4,5], but we should be clear that their approach is flawed. This is because they include compounds in their correspondence values. As mentioned above in the discussion with regard to Figs. 2 and 4, we need only focus on reactions (since for each reaction we know the set of compounds involved). As many reactions involve more than one input/output compound, the correspondence measures used by Croes et al. [4,5] could classify a computed path as less than perfect, even if it contains exactly the same set of reactions as the metabolic path (due to different compounds being involved in the metabolic path and the computed path). For example, referring to Fig. 4, suppose the metabolic path is C1 → R1 → C2 → R2 → C5 → R3 → C7 (the left-hand path shown in Fig. 4), but the computed path is C1 → R1 → C3 → R2 → C5 → R3 → C7 (the right-hand path shown in Fig. 4). This will give TP = 6, FP = 1, FN = 1, Sn = 6/7, PPV = 6/7, Ac = 6/7. Yet both paths contain precisely the same reactions, which is the key feature. Clearly the correspondence values defined by Croes et al. [4,5] are inappropriate. Better correspondence measures would drop all mention of compounds in the values defined above. Despite this flaw we, for reasons of consistency of comparison with the results presented previously in Croes et al. [4,5], will present our results below using the correspondence measures that include compounds as defined above.

3.2. Results for an example known pathway—Gluconeogenesis

Table 1 shows the correspondence values for each of the k-shortest paths ($k = 1, 2, \ldots, 10$) computed from the initial reaction to final reaction (the R–R case) in the Gluconeogenesis pathway. For $k = 1$, i.e. the shortest path, there is a low level of correspondence between the metabolic path and the computed shortest path. For $k = 2$, i.e. the second shortest path, correspondence increases (sensitivity, positive predictive value and accuracy all increase). Note though, that as we increase k we find different paths, and so there is no guarantee that correspondence increases with increasing k. For $k = 4$, for example, the correspondence values decrease—so the 4-shortest path corresponds less well to the metabolic path than the 3-shortest path. In fact the correspondence between the 4-shortest path and the metabolic path is less than for the ($k = 1$) shortest path. It can be seen from Table 1 that for $k = 6$ the solution is precisely the same as the Gluconeogenesis metabolic path.

Table 3
Values for the best correspondence path among the first k-shortest paths for $k = 1, 5, 10$ in the Gluconeogenesis pathway

Case	k	Sensitivity (Sn)	Positive predictive value (PPV)	Accuracy (Ac)
R–R	1	0.385	0.714	0.549
	5	0.769	0.909	0.839
	10	1	1	1
C–C	1	0.133	0.118	0.125
	5	0.467	0.778	0.622
	10	0.467	0.778	0.622

Table 4
Pathways examined

Pathway number	Pathway name
1	Gluconeogenesis
2	Glycogen
3	Glycolysis
4	Proline biosynthesis
5a	Ketogluconate metabolism
5b	Ketogluconate metabolism
6a	Pentose phosphate
6b	Pentose phosphate
7	Salvage pathway deoxythymidine phosphate
8	Tricarboxylic acid (citric acid, citrate, TCA, Krebs) cycle
9	NAD biosynthesis
10a	Arginine biosynthesis
10b	Arginine biosynthesis

Table 2 shows the correspondence values for each of the k-shortest paths ($k = 1, 2, \ldots, 10$) computed from the source compound to the target compound (the C–C case) in the Gluconeogenesis pathway. It can be seen that correspondence is markedly less than for the R–R case, and for no value of k examined is the computed k-shortest path the same as the Gluconeogenesis metabolic path.

Table 3 shows correspondence values for the best correspondence path (as measured by maximum accuracy) amongst all of the first k-shortest paths for a number of different values of k for the Gluconeogenesis pathway. For $k = 5$ in the C–C case, for example, the best correspondence path out of the first five shortest paths has accuracy 0.622. Examining Table 2, we can see that this path was the second shortest path. Because here we take the maximum accuracy path from amongst the first k-shortest paths correspondence increases as we increase k.

3.3. Results for ten known pathways—correspondence values

In this paper, we have examined the ten pathways (including Gluconeogenesis) shown in Table 4. One complication arose with pathways 5 (Ketogluconate metabolism), 6 (Pentose phosphate) and 10 (Arginine biosynthesis) in the R–R case in that the definition of the first or last reaction turned out to be ambiguous, there being two different options for the first reaction (pathways 5 and 10) or for the last reaction (pathway 6), as can be seen in the accompanying supplement. Consequently, we computed two different metabolic paths in the R–R case for each pathway.

As far as the C–C case is concerned, one minor issue relates to pathway 8, the TCA cycle. In this pathway, the source compound and the target compound are the same. The usual definition of a path is that the initial and final nodes are different (whereas in a cycle the initial and final nodes are the same). Hence in order to deal with this pathway, we treated the source/target compound as two different compounds, one relating to being used as input to a reaction, the other relating to being used as output from a reaction.

Detailed results for all of the pathways shown in Table 4 can be found in the accompanying supplement.

Table 5 shows the same information as Table 3, but averaged over all the pathways considered. In Table 5 we, for example, have that the average accuracy for the shortest path ($k = 1$) is 0.843 in the R–R case, but only 0.449 in the C–C case. If we take, for each of the pathways examined, the maximum accuracy (best correspondence) path over the first five shortest paths this accuracy increases to 0.969 in the R–R case and 0.818 in the C–C case.

Table 5 indicates high correspondence for the R–R case. These results are in accordance with the results presented in Croes et al. [4,5]. However, one of the deficiencies of the Croes et al. [4,5] work is that no results are presented for the C–C case. This is especially important as, in the literature, metabolic pathways are typically viewed as relating to transforming one compound into another—not as relating to going from one reaction to another. It is clear from Table 5 that correspondence is poor for the C–C case. Even taking the first ten shortest paths correspondence (average maximum accuracy) is only 0.818—less that the correspondence achieved for the shortest path ($k = 1$) in the R–R case.

Poor maximum accuracy in the C–C case is especially found in those metabolic paths, where the number of intermediate reactions/compounds between the source compound and the target compound is high. Fig. 6 plots, for each pathway in the

Table 5
Values for the best correspondence path among the first k-shortest paths for $k = 1, 5, 10$ averaged over all pathways

Case	k	Sensitivity (Sn)	Positive predictive value (PPV)	Accuracy (Ac)
R–R	1	0.828	0.858	0.843
	5	0.953	0.984	0.969
	10	0.970	0.991	0.981
C–C	1	0.400	0.499	0.449
	5	0.755	0.882	0.818
	10	0.755	0.882	0.818

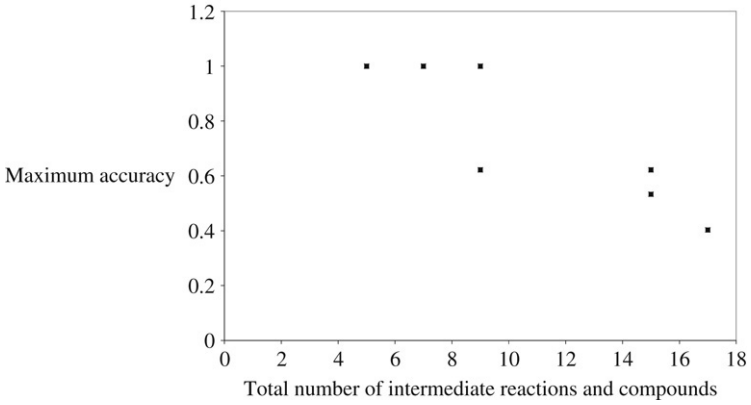

Fig. 6. Maximum accuracy over all ten k-shortest paths for each pathway as against number of intermediate reactions/compounds for the C–C case.

Table 6
Metabolic path recovery amongst the first k-shortest paths for $k = 1, 5, 10$ for the R–R case

Pathway number	Pathway name	Metabolic path recovered?		
		$k = 1$	$k = 5$	$k = 10$
1	Gluconeogenesis	No	No	Yes
2	Glycogen	Yes	Yes	Yes
3	Glycolysis	No	Yes	Yes
4	Proline biosynthesis	Yes	Yes	Yes
5a	Ketogluconate metabolism	Yes	Yes	Yes
5b	Ketogluconate metabolism	Yes	Yes	Yes
6a	Pentose phosphate	Yes	Yes	Yes
6b	Pentose phosphate	Yes	Yes	Yes
7	Salvage pathway deoxythymidine phosphate	No	Yes	Yes
8	Tricarboxylic acid (citric acid, citrate, TCA, Krebs) cycle	No	No	No
9	NAD biosynthesis	Yes	Yes	Yes
10a	Arginine biosynthesis	Yes	Yes	Yes
10b	Arginine biosynthesis	Yes	Yes	Yes
Number of "yes" entries (maximum 13)		9	11	12

C–C case, the maximum accuracy over all ten shortest paths against the number of intermediate reactions/compounds in the metabolic path. As can be seen from Fig. 6, maximum accuracy declines as the metabolic path involves more intermediate reactions/compounds. One further point to be made from Table 5 is that the results are not significantly improved in either the R–R or C–C cases by moving from the first five shortest paths to the first ten shortest paths. In other words, the computation of more shortest paths beyond the first five is of little (average) benefit.

3.4. Results for ten known pathways—metabolic path recovery

Whilst Table 5 gives an insight into accuracy, we believe that it is appropriate to also tabulate whether, or not, a computed shortest path corresponds *exactly* to the metabolic path (which we term *"recovering"* the path). Clearly, recovering a metabolic path is the ideal case (and corresponds to an accuracy (Ac) of one). Table 6 indicates for the R–R case whether, or not, we recover the metabolic path amongst the first k-shortest paths for $k = 1, 5, 10$.

Table 7 presents the same information as Table 6 but for the C–C case. It also shows the results given in Beasley and Planes [3], where results are presented for two objectives (labelled 13 and 14) in terms of their success at recovering metabolic pathways.

Table 7
Metabolic path recovery amongst the first k-shortest paths for $k = 1, 5, 10$ for the C–C case and comparison with Beasley and Planes [3]

Pathway number	Pathway name	Metabolic path recovered?			Beasley and Planes [3]	
		$k = 1$	$k = 5$	$k = 10$	Metabolic pathway recovered?	
					Objective (13)	Objective (14)
1	Gluconeogenesis	No	No	No	Yes	No
2	Glycogen	Yes	Yes	Yes	Yes	No
3	Glycolysis	No	No	No	Yes	Yes
4	Proline biosynthesis	No	No	No	Yes	No
5	Ketogluconate metabolism	Yes	Yes	Yes	No	No
6	Pentose phosphate	No	Yes	Yes	Yes	No
7	Salvage pathway deoxythymidine phosphate	No	Yes	Yes	Yes	No
8	Tricarboxylic acid (citric acid, citrate, TCA, Krebs) cycle	No	No	No	No	Yes
9	NAD biosynthesis	No	Yes	Yes	Yes	No
10	Arginine biosynthesis	Yes	Yes	Yes	Yes	No
Number of "yes" entries (maximum 10)		3	6	6	8	2

Table 6 indicates that for the R–R case, the ($k = 1$) shortest path recovers the metabolic path in 9 out of 13 pathways—this figure rising to recovering 12 of the 13 paths if we consider $k = 10$. On the other hand, Table 7 indicates that for the C–C case the ($k = 1$) shortest path recovers the metabolic path in only 3 out of 10 pathways—this figure rising to recovering 6 of the 10 paths if we consider $k = 10$.

Comparing the results shown in Table 7 ($k = 10$) with the results for Beasley and Planes [3] (taking the best of both objectives), we have a mix of situations: some where both approaches achieve recovery (e.g. pathway 2); some where the path finding approach presented here achieves recovery and Beasley and Planes [3] do not (e.g. pathway 5); and some where the path finding approach presented here does not achieve recovery and Beasley and Planes [3] do (e.g. pathway 1).

3.5. Discussion

In essence, the path finding approach to metabolic pathways given above rests on the hypothesis that insight into a metabolic pathway can be obtained by finding k-shortest paths, using compound node connectivities as a distance metric. Our results partially support this hypothesis.

It is clear that for the R–R case, where our results are in accordance with the results presented previously in Croes et al. [4,5], this hypothesis is valid. We have high correspondence values, and recover the metabolic path for nearly all pathways examined. However for the C–C case the validity of the hypothesis is more questionable. Correspondence values are not as good as for the R–R case and we recover far fewer metabolic paths.

Clearly, the lack of success for the C–C case, as opposed to the R–R case, could be due to a number of factors. It may be that k-shortest paths (whatever the distance metric used) is not an appropriate concept for analysing metabolic pathways. It is clear that the literature is divided as to whether, or not, making use of shortest paths is of value with respect to metabolic pathways. Some workers (e.g. Mavrovouniotis [14–17], Schuster et al. [26]) do not use shortest paths, others (e.g. Arita [2], Croes [4,5]) do. However, if utilising shortest paths were inappropriate, then we would not have expected the results for the R–R case to be any better than the results for the C–C case. But in fact we find that the results for the R–R case are better than the results for the C–C case.

It could be, of course, that k-shortest paths are an appropriate concept for analysing metabolic pathways, but we have used an inappropriate distance metric. The distance metric used in this paper related to the connectivity of the compounds involved in the path. Other possibilities (obviously) exist. For example, we might use a distance metric based on reactions. Such a metric, for example, could be related to the amount of chemical change that takes place at each reaction, or to energetic considerations such as the Gibbs free energy for each reaction. Alternatively, a distance metric that takes both compounds and reactions into account may be appropriate.

Finally, we would note here, that the key difference between the C–C case and the R–R case relates to the fact that in the C–C case we have to choose two extra reactions in the path: one reaction having the source compound as an input compound, the other reaction having the target compound as an output compound. In the R–R case, these two reactions are specified. This might imply that taking the current distance metric (which is compound based, but which is successful for the R–R case), and amending it for the C–C case with reaction terms that relate only to any reactions having the source compound as an input compound, or having the target compound as an output compound, could be a profitable approach.

4. Conclusions

In this paper, the effectiveness of using compound node connectivities in a path finding approach to metabolic pathways has been examined. We found that finding k-shortest paths using a distance metric based on compound node connectivities performed well, when a metabolic path was regarded as being from a source reaction to a target reaction. The same approach performed less well when a metabolic path was regarded as being from a source compound to a target compound. An

approach to path finding based upon integer programming was also presented that produces k-shortest paths in increasing distance order, and guarantees that the paths found will be optimal.

Acknowledgements

The authors would like to acknowledge the detailed comments provided by one of the anonymous referees.

Appendix. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.dam.2008.06.035.

References

[1] T. Aittokallio, B. Schwikowski, Graph-based methods for analysing networks in cell biology, Briefings in Bioinformatics 7 (2006) 243–255.
[2] M. Arita, Metabolic reconstruction using shortest paths, Simulation Practice and Theory 8 (2000) 109–125.
[3] J.E. Beasley, F.J. Planes, Recovering metabolic pathways via optimisation, Bioinformatics 23 (2007) 92–98.
[4] D. Croes, F. Couche, S.J. Wodak, J. van Helden, Metabolic PathFinding: Inferring relevant pathways in biochemical networks, Nucleic Acids Research 33 (2005) W326–W330.
[5] D. Croes, F. Couche, S.J. Wodak, J. van Helden, Inferring meaningful pathways in weighted biochemical networks, Journal of Molecular Biology 356 (2006) 222–236.
[6] G. Dooms, Y. Deville, P. Dupont, Constrained metabolic network analysis: Discovering pathways using CP(Graph), 2005. http://www2.info.ucl.ac.be/people/YDE/Papers/wcb05.pdf.
[7] F. Guerriero, R. Musmanno, V. Lacagnina, A. Pecorella, A class of label-correcting methods for the K shortest paths problem, Operations Research 49 (2001) 423–429.
[8] V. Hatzimanikatis, C. Li, J.A. Ionita, C.S. Henry, M.D. Jankowski, L.J. Broadbelt, Exploring the diversity of complex metabolic networks, Bioinformatics 21 (2005) 1603–1609.
[9] ILOG CPLEX, 2005. http://www.ilog.com/products/cplex/news/whatsnew.cfm#cplex90.
[10] I.M. Keseler, J. Collado-Vides, S. Gama-Castro, J. Ingraham, S. Paley, I.T. Paulsen, M. Peralta-Gill, P.D. Karp, EcoCyc: A comprehensive database resource for Escherichia coli, Nucleic Acids Research 33 (2005) D334–D337.
[11] S. Klamt, J. Stelling, Two approaches for metabolic pathway analysis?, Trends in Biotechnology 21 (2003) 64–69.
[12] I. Koch, B.H. Junker, M. Heiner, Application of Petri net theory for modelling and validation of the sucrose breakdown pathway in the potato tuber, Bioinformatics 21 (2005) 1219–1226.
[13] R. Küffner, R. Zimmer, T. Lengauer, Pathway analysis in metabolic databases via differential metabolic display (DMD), Bioinformatics 16 (2000) 825–836.
[14] M.L. Mavrovouniotis, G. Stephanopoulos, G. Stephanopoulos, Computer-aided synthesis of biochemical pathways, Biotechnology and Bioengineering 36 (1990) 1119–1132.
[15] M.L. Mavrovouniotis, Synthesis of reaction mechanisms consisting of reversible and irreversible steps. 1. A synthesis approach in the context of simple examples, Industrial and Engineering Chemistry Research 31 (1992) 1625–1637.
[16] M.L. Mavrovouniotis, Synthesis of reaction mechanisms consisting of reversible and irreversible steps. 2. Formalization and analysis of the synthesis algorithm, Industrial and Engineering Chemistry Research 31 (1992) 1637–1653.
[17] M.L. Mavrovouniotis, Identification of qualitatively feasible metabolic pathways, in: L. Hunter (Ed.), Artificial Intelligence and Molecular Biology, AAAI Press/MIT Press, 1993, pp. 325–364.
[18] D.C. McShan, S. Rao, I. Shah, PathMiner: Predicting metabolic pathways by heuristic search, Bioinformatics 19 (2003) 1692–1698.
[19] D.L. Nelson, M.M. Cox, Lehninger Principles of Biochemistry, fourth edn, Worth Publishers, New York, 2005.
[20] J.A. Papin, N.D. Price, S.J. Wiback, D.A. Fell, B.O. Palsson, Metabolic pathways in the post-genome era, Trends in Biochemical Science 28 (2003) 250–258.
[21] S.A. Rahman, P. Advani, R. Schunk, R. Schrader, D. Schomburg, Metabolic pathway analysis web service (Pathway Hunter Tool at CUBIC), Bioinformatics 21 (2005) 1189–1193.
[22] S.A. Rahman, D. Schomburg, Observing local and global properties of metabolic pathways: 'Load points' and 'choke points' in the metabolic networks, Bioinformatics 22 (2006) 1767–1774.
[23] J.L. Reed, T.D. Vo, C.H. Schilling, B.O. Palsson, An expanded genome-scale model of Escherichia coli K-12 (*i*JR904 GSM/GPR), Genome Biology 4 (2003) R54.
[24] C.H. Schilling, D. Letscher, B.O. Palsson, Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective, Journal of Theoretical Biology 203 (2000) 229–248.
[25] C.H. Schilling, S. Schuster, B.O. Palsson, R. Heinrich, Metabolic pathway analysis: Basic concepts and scientific applications in the post-genomic era, Biotechnology Progress 15 (1999) 296–303.
[26] S. Schuster, D.A. Fell, D. Dandekar, A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks, Nature Biotechnology 18 (2000) 326–332.
[27] A. Seressiotis, J.E. Bailey, MPS: An algorithm and data base for metabolic pathway synthesis, Biotechnology Letters 8 (1986) 837–842.
[28] A. Seressiotis, J.E. Bailey, MPS—An artificially intelligent software system for the analysis and synthesis of metabolic pathways, Biotechnology and Bioengineering 31 (1988) 587–602.
[29] E. Simao, E. Remy, D. Thieffry, C. Chaouiya, Qualitative modelling of regulated metabolic pathways: Application to the tryptophan biosynthesis in *E. Coli*, Bioinformatics 21 (Suppl. 2) (2005) ii190–ii196.
[30] J. Stelling, S. Klamt, K. Bettenbrock, S. Schuster, E.D. Gilles, Metabolic network structure determines key aspects of functionality and regulation, Nature 420 (2002) 190–193.
[31] I. Zevedei-Oancea, S. Schuster, Topological analysis of metabolic networks based on Petri net theory, In Silico Biology 3 (2003) 0029. http://www.bioinfo.de/isb/2003/03/0029/main.html.